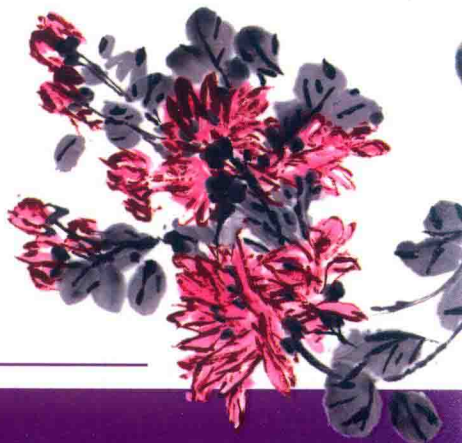


21世纪高等院校通识教育规划教材



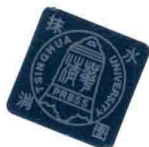
DAXUESHENG XINXI JIANSUO SUYANG JIAOCHENG

大学生信息检索素养教程

王冲 编著
Wang Chong



2



清华大学出版社

21世纪高等院校通识教育规划教材



DAXUESHENG XINXI JIANSUO SUYANG JIAOCHENG

大学生信息检索素养教程

王冲 编著
Wang Chong

清华大学出版社
北京

内 容 简 介

本书属于高等学校各个专业研究生和本科生的“信息检索素养课程”教学通用教材,内容包括三大部分:第一部分“信息检索素养基础知识篇”,第二部分“信息检索素养基本原理篇”和第三部分“信息检索素养实践应用篇”,共13章内容。本书较好地现代信息检索素养知识的基础性与前沿性、原理性与实践性、全面性与主题性、引导性与启发性进行了贯通与融合。在基于大量信息检索专题、图表、实例及其数学理论依据进行充分阐述和说明的基础上,突出国内与国外、理论与实践紧密结合的信息检索素养教学要求。考虑到不同专业和不同层次学生的实际教学需要,教学内容组织依据循序渐进和主题性教学相结合的原则,可以适当选用部分章节组织教学。例如,针对计算机学科专业、图书情报学专业、信息管理专业本科生和各个专业的研究生层次学生,可以把第二部分“信息检索素养基本原理篇”作为重点来组织各个教学章节内容。

本书内容丰富、线索清晰、结构完整、语言精练、主题鲜明,是高等学校各个专业研究生和本科生的信息检索素养教学通用教材。既可以作为信息检索素养基础必修课教材,也可以作为部分专业和图书馆用户教育的选修课教材,同时可作为信息系统设计与开发、数据采集与挖掘、信息检索与咨询服务、图书情报机构等从业人员的学习与培训参考用书。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。
版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

大学生信息检索素养教程 / 王冲编著. —北京:清华大学出版社, 2017
(21世纪高等院校通识教育规划教材)
ISBN 978-7-302-46006-0

I. ①大… II. ①王… III. ①信息检索—高等学校—教材 IV. ①G254.9

中国版本图书馆CIP数据核字(2016)第313719号

责任编辑:白立军 薛 阳

封面设计:傅瑞学

责任校对:梁 毅

责任印制:杨 艳

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦A座

社总机:010-62770175

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

课件下载: <http://www.tup.com.cn>, 010-62795954

邮 编: 100084

邮 购: 010-62786544

印 刷 者: 北京富博印刷有限公司

装 订 者: 北京市密云县京文制本装订厂

经 销: 全国新华书店

开 本: 185mm×230mm

印 张: 34.75

字 数: 674千字

版 次: 2017年1月第1版

印 次: 2017年1月第1次印刷

印 数: 1~2000

定 价: 59.50元

产品编号: 059891-01

前言

在信息化社会越来越发达的今天,面对几何级数膨胀的海量信息资源,如何有效地检索、获取、评估、传播、共享和利用信息,成为了每个人重要的基本素养和能力要求,因为信息需求是每个人学习、工作、生活及其社会活动中十分重要而且迫切的需求。作为信息时代的大学生,需要重视信息检索素养的知识学习与能力培养。信息检索素养的理论知识学习与基本能力形成,不仅直接影响着大学生的在校学业表现,也较大程度上影响着他们今后的学习、工作与事业发展(例如终身学习、创新创业等持续性需要)。

大学生信息检索素养是大学生信息素养的核心内容之一,具有多学科交叉融合的特性。信息检索起源于图书馆学、情报学的信息检索原理与技术,早期直接服务于高校图书馆或社会公共图书馆的信息检索用户教育与技能培训,后来广泛应用于数据库研发与服务企业、搜索引擎等信息服务产业,在当今高速发展的计算机科学、软件工程、网络工程、通信工程、管理学、应用数学、统计学、语言学等多学科交叉融合的基础上,信息检索在数据挖掘、大数据处理等领域不断深化并发挥着日益强大的潜能。大学生信息检索素养教育正是基于这种时代背景和学科发展提出来的,也是面向大学生的传统信息素养教育和信息检索教育的不断深化与交叉融合的发展结果。

基于循序渐进和主题性教学原则,本书较好地把握现代信息检索素养知识的原理性与实践性、全面性与主题性、引导性与启发性进行了贯通与融合。在基于大量信息检索原理与知识的专题、图表、实例、案例及其数学理论依据进行充分阐述和说明的基础上,突出国内与国外、基础与前瞻、知识与技能紧密结合的信息检索素养教学要求。考虑到不同专业和不同层次学生的实际教学需要,本教材属于高等学校各个专业研究生和本科生的“信息检索素养课程”通用教材,内容包括三大部分:第一部分“信息检索素养基础知识篇”,第二部分“信息检索素养基本原理篇”和第三部分“信息检索素养实践应用篇”。

本书逻辑清晰,内容丰富,结构完整。首先,从信息检索素养的基本概念、内涵、发展动因、特点、核心内容与能力表现、信息检索素养的评价标准以及信息化社会对大学生的信息检索素质需要出发,进一步论述信息检索与知识产权、信息检索与大学生学术不端行为、信息检索基础知识、信息检索方法与策略等内容来培养学生的信息检索意识、信息检

索道德与信息检索基础。第二,通过“信息检索的基础数学原理”的引入,使得信息检索有了更加严谨的逻辑论证,检索过程和信息需求的本质描述也更为精确,从而使得信息检索的理论与实践获得持续性的基础支撑。通过“文本分类与文本索引构建”、“图像信息检索”、“音频信息检索”、“视频信息检索”和“Web 信息搜索一般性原理”来构建大学生特别是研究生的信息检索基本原理知识。第三,通过“搜索引擎的检索应用”、“七大类特种文献信息资源检索”和“图书与学术期刊论文检索”的大量实例与检索案例来培养和锻炼大学生的信息检索素养实践技能。

本书教学内容的规划、组织与编著,是在作者讲授研究生“信息检索原理与应用”课程和本科生“大学生信息检索”课程的十多年教学改革与实践经验基础上逐步积累形成的。同时,在教材编著过程中,参考和借鉴了大量国内外专著、教材、学术期刊论文、学位论文、学术观点和典型网络数据库检索平台等成果,在此一并向他们表示真挚的谢意!

本书内容丰富、线索清晰、结构完整、语言精练、主题鲜明,是高等学校各个专业研究生和本科生的信息检索素养教学通用教材。既可以作为信息检索素养基础必修课教材,也可以作为部分专业和图书馆用户教育的选修课教材,同时可作为信息系统设计与开发、数据采集与挖掘、信息检索与咨询服务、图书情报机构等从业人员的学习与培训参考用书。

在本书编著过程中,得到桂林电子科技大学研究生院领导及教学督导委员会的关心与支持,获得“2016年桂林电子科技大学研究生教育质量工程专项(YXYJ2900)”、“2016年广西学位与研究生教育改革与发展专项(2016XWYJ12)”和“2015年广西高等教育本科教学改革工程项目(2015JGA207)”的支持与资助。本书能够顺利出版,感谢清华大学出版社的大力支持与良好合作,感谢出版社编辑们的辛勤工作与付出!

本书主要基于循序渐进性教学与主题性教学相结合的编写原则,在大学生信息检索素养的原理性与实践性、全面性与主题性、引导性与启发性等方面难免有疏漏或不妥之处,恳请读者批评指正。

作者

2016年7月于桂林

目 录

第一部分 信息检索素养基础知识篇

第1章 大学生信息检索素养概述	3
1.1 信息检索素养概述	4
1.1.1 信息检索素养的基本概念	4
1.1.2 大学生信息检索素养的内涵	5
1.1.3 信息检索素养的发展动因	6
1.1.4 信息检索素养的特点	7
1.2 信息检索素养的主要内容	9
1.2.1 信息检索意识	9
1.2.2 信息检索能力	10
1.2.3 信息检索道德	10
1.3 信息检索素养的评价标准	11
1.3.1 有信息检索素养的人	11
1.3.2 信息检索素养评价标准的必要性	12
1.3.3 大学生信息检索素养评价标准	13
1.4 我国当代大学生的信息检索素养现状	14
1.4.1 信息检索意识较弱	14
1.4.2 获取信息的检索能力不强	14
1.4.3 加工与利用信息的能力较差	14
1.4.4 信息道德和信息法规意识急需培养	14
1.5 大学生信息检索素养教育与培养的意义	15
1.5.1 信息化社会对大学生的信息检索素质需求	15
1.5.2 创新创业能力培养的需要	16
1.5.3 掌握有效信息和开展科研与学术活动的需要	17

1.5.4 提供科学方法与正确决策的需要	18
1.5.5 终身学习的需要	19
本章小结	19
本章思考与练习题	21
第2章 信息检索与知识产权	22
2.1 信息与知识产权	22
2.1.1 信息	22
2.1.2 知识产权	26
2.1.3 知识产权信息	27
2.1.4 知识产权信息的概念特征	28
2.1.5 知识产权信息的内容	29
2.2 信息检索与利用的法律规范和信息道德	29
2.2.1 信息检索与利用的相关法律制度	30
2.2.2 知情权问题	31
2.2.3 国家秘密问题	32
2.2.4 商业秘密问题	33
2.2.5 隐私权保护问题	33
2.2.6 信息复制权保护问题	34
2.3 信息检索与利用过程中的道德自律	34
2.3.1 法律约束的局限性	35
2.3.2 信息道德自律问题的提出	35
2.3.3 信息道德的培养和内省原则	36
2.4 信息检索与利用同知识产权保护的影响	36
2.4.1 信息检索与利用对知识产权保护既制约又促进	36
2.4.2 知识产权保护对信息检索与信息资源共享的制约和促进	37
2.5 大学生信息检索素养与学术不端行为的关联	38
2.5.1 大学生学术不端行为的界定	38
2.5.2 大学生学术不端行为的表现	39
2.5.3 信息检索素养教育对大学生学术不端行为的作用	40
本章小结	41

本章思考与练习题	43
第 3 章 信息检索的基本知识	44
3.1 信息检索的含义	44
3.1.1 检索的概念	44
3.1.2 信息检索的含义	45
3.1.3 信息检索用户的基础素养	46
3.1.4 信息检索的领域与范畴	47
3.1.5 信息检索的类型	48
3.2 信息检索涉及的相关支撑领域	49
3.3 信息检索的前沿与热点问题	51
3.3.1 信息检索的发展趋势	51
3.3.2 信息检索的热点问题	55
本章小结	57
本章思考与练习题	58
第 4 章 信息检索的方法与策略	59
4.1 信息源及其类型	59
4.2 信息源的出版发行与共享类型	61
4.3 信息源类型的辨别	64
4.4 检索工具	67
4.4.1 检索工具的基本功能	67
4.4.2 检索工具的类型	69
4.5 信息检索途径	73
4.6 信息检索方法	82
4.7 信息检索策略	84
4.8 信息检索质量与评价	87
4.8.1 信息检索质量与评价指标	88
4.8.2 影响检索效果的因素	89
本章小结	91

本章思考与练习题	91
----------------	----

第二部分 信息检索素养基本原理篇

第5章 信息检索的基础数学原理	95
5.1 简单布尔检索	95
5.1.1 基本原理	95
5.1.2 布尔检索模型的特点	97
5.2 信息检索模糊集合论	98
5.2.1 模糊检索的数学描述	99
5.2.2 信息文档对标引词的隶属度	100
5.2.3 提问检索词的相关性描述	100
5.3 扩展布尔检索	102
5.3.1 基于两个标引词的情形	102
5.3.2 推广到 n 个标引词空间	103
5.4 信息检索代数模型	106
5.4.1 信息检索向量空间模型	106
5.4.2 潜在语义索引模型	113
5.4.3 神经网络检索模型	117
5.5 概率论检索模型	122
5.5.1 经典概率检索模型	123
5.5.2 贝叶斯网络检索模型	125
5.6 其他检索模型的一般数学原理	129
5.6.1 进化计算与遗传算法	129
5.6.2 粗糙集理论	136
5.6.3 浏览检索模型	140
本章小结	142
本章思考与练习题	144
第6章 文本分类与文本索引构建	145
6.1 文本分类概述	146

6.2	朴素贝叶斯文本分类	148
6.2.1	贝叶斯分类器	148
6.2.2	条件概率和乘法定理	149
6.2.3	极大后验假设和极大似然假设	149
6.2.4	贝叶斯定理	150
6.2.5	多项式朴素贝叶斯	151
6.3	朴素贝叶斯分类模型改进	153
6.3.1	改进方法	153
6.3.2	朴素贝叶斯分类的提升模型	155
6.3.3	基于特征相关的改进加权朴素贝叶斯分类	156
6.4	贝努利文本分类模型	157
6.5	多项式文本分类模型与贝努利文本分类模型的性质比较	159
6.6	文本分类特征选择	161
6.6.1	文本分类特征选择的作用	161
6.6.2	特征选择的方法	162
6.6.3	特征选择方法类型	163
6.6.4	文本互信息选择	164
6.6.5	χ^2 统计量特征选择	165
6.6.6	基于频率的特征选择方法	166
6.7	文本的索引构建	167
6.7.1	基于块的排序索引方法	167
6.7.2	基于内存单次扫描的索引构建方法	171
6.7.3	顺排文档索引	172
6.7.4	倒排文档索引	178
	本章小结	186
	本章思考与练习题	187
第 7 章	图像信息检索	189
7.1	图像基础知识	189
7.1.1	图像色彩三要素	190
7.1.2	图像的三种基本类型	192

7.1.3	常用图像文件格式	192
7.2	图像检索概述	196
7.2.1	图像检索一般模型	196
7.2.2	基于文本方式的图像检索	197
7.2.3	基于知识和视觉特征的图像检索	198
7.2.4	基于内容的图像检索	198
7.2.5	图像内容描述的标准化	199
7.3	基于图像内容特征提取	200
7.3.1	基于颜色特征的图像检索	200
7.3.2	基于纹理特征的图像检索	204
7.3.3	基于形状特征的图像检索	206
7.3.4	基于空间特征的图像检索	214
7.3.5	单个特征图像检索的不足	215
7.4	基于多特征的图像检索	216
7.4.1	综合颜色和形状特征的图像检索	216
7.4.2	综合形状和空间特征的图像检索	216
7.4.3	综合形状和纹理特征的图像检索	217
7.4.4	综合颜色、形状和空间的图像检索	217
7.5	基于视觉特征的图像检索系统	218
7.5.1	基于视觉特征的图像检索系统整体架构	218
7.5.2	图像分割技术	219
7.5.3	相似性度量	224
7.5.4	图像索引	226
7.5.5	相关反馈技术	232
7.6	典型的图像检索系统	233
7.7	图像检索技术的发展方向	234
7.7.1	融合人工反馈	234
7.7.2	高层语义和低层视觉特征结合	234
7.7.3	面向网络图像检索	235
7.7.4	图像检索性能评价与检索服务平台	235
	本章小结	236

本章思考与练习题	237
第 8 章 音频信息检索	239
8.1 音频的特点	239
8.1.1 音频信息的基本特征	239
8.1.2 音频信息的内容层次	240
8.2 音频信息检索技术的分类和发展	241
8.2.1 基于文本的音频检索	241
8.2.2 基于内容特征的音频检索	243
8.3 音频信息检索架构与模型	244
8.3.1 音频信息检索架构	244
8.3.2 向量空间模型借鉴	245
8.3.3 概率模型借鉴	246
8.4 表示级的音频检索	247
8.4.1 基于直接匹配的音频样例检索	247
8.4.2 基于索引的音频样例检索	249
8.4.3 基于 GPU 通用计算的音频样例快速检索	256
8.5 语义级的语音文档检索	263
8.5.1 语音文档检索的预处理	263
8.5.2 语音文档检索的索引和搜索技术	266
8.5.3 语音文档检索中的容错方法	270
本章小结	274
本章思考与练习题	275
第 9 章 视频信息检索	277
9.1 数字视频的相关基础知识	277
9.2 基于内容的视频检索系统结构	280
9.3 视频镜头分割	281
9.3.1 非压缩域的镜头分割方法	282
9.3.2 压缩域中镜头分割方法	285
9.4 镜头切换	286

9.5	关键帧提取及语义提取	287
9.5.1	关键帧提取的基本原理和准则	287
9.5.2	关键帧提取的方法	287
9.5.3	视频语义提取	290
9.6	视频特征提取	291
9.6.1	全局运动矢量的计算方法	292
9.6.2	视频运动估计	293
9.6.3	运动矢量估计的常用算法	296
9.7	视频聚类	301
9.8	视频结构索引	302
9.8.1	视频结构索引的机制	303
9.8.2	索引信息的存储	303
9.9	视频摘要	305
9.10	视频语义检索模型	308
9.10.1	底层特征提取模块	308
9.10.2	底层特征向高层语义映射模块	308
9.10.3	视频语义查询模块	310
9.10.4	语义词典的应用	311
9.11	典型的视频检索系统	311
	本章小结	312
	本章思考与练习题	314
第 10 章	Web 信息搜索	316
10.1	搜索引擎概述	316
10.1.1	搜索引擎基本结构	317
10.1.2	传统搜索引擎基本类型	318
10.1.3	智能搜索引擎基本类型	319
10.2	搜索引擎主要支撑技术	324
10.2.1	分词技术	324
10.2.2	网络蜘蛛	325
10.2.3	索引技术	325

10.2.4	词频相关指数	326
10.2.5	自动推理技术	326
10.2.6	本体知识系统	327
10.2.7	专家系统	328
10.3	Web 采集	329
10.3.1	Web 采集概述	329
10.3.2	采集器的功能与特点	329
10.3.3	Web 采集	330
10.3.4	域名解析	332
10.3.5	待采集 URL 池	335
10.3.6	分布式索引	336
10.3.7	连接服务器	339
10.3.8	Web 图	340
10.4	主要网页排序算法	342
10.4.1	PageRank 网页排序算法	343
10.4.2	Topic-Sensitive PageRank 算法	343
10.4.3	Hilltop 算法	344
10.4.4	HITS 算法	345
10.4.5	SALSA 算法	346
10.4.6	BFS 算法	347
10.4.7	PHITS 算法	347
	本章小结	348
	本章思考与练习题	349

第三部分 信息检索素养实践应用篇

第 11 章	常用搜索引擎的检索应用	353
11.1	百度搜索引擎的检索应用	353
11.2	搜狗搜索引擎的信息检索与利用	372
11.3	Google 搜索引擎的检索应用	384
11.4	Infoseek 搜索引擎	392

11.5	雅虎搜索引擎信息检索应用	396
	本章小结	399
	本章思考与练习题	400
第 12 章	特种信息资源检索	401
12.1	科技报告信息资源检索	401
12.1.1	科技报告的概念与特征	401
12.1.2	科技报告的类型与编码	402
12.1.3	国内科技报告与商业报告资源的信息检索	403
12.1.4	国外科技报告资源检索	409
12.2	会议文献资源检索	413
12.2.1	会议文献资源的概念	413
12.2.2	会议文献的特点与类型	414
12.2.3	国外会议文献的检索	415
12.2.4	国内会议文献的检索	419
12.3	学位论文检索	423
12.3.1	学位论文概述	423
12.3.2	国外重要学位论文数据库检索	424
12.3.3	重要国内学位论文数据库检索	426
12.4	专利文献资源检索	434
12.4.1	专利与专利文献概念	434
12.4.2	专利文献的类型与作用	434
12.4.3	国际专利分类	436
12.4.4	专利搜索引擎	438
12.4.5	国外大型专利数据库系统	445
12.4.6	国内专利资源数据库系统检索	455
12.5	标准信息资源检索	462
12.5.1	标准信息资源的概念与特点	462
12.5.2	标准信息资源的分类	463
12.5.3	美英等国标准信息资源检索	464
12.5.4	中文标准信息资源检索	467

本章小结	471
本章思考与练习题	472
第 13 章 图书与学术期刊论文信息资源检索	474
13.1 大型中文图书目录检索系统	474
13.1.1 中国国家图书馆联机公共目录查询系统	474
13.1.2 CALIS 联合目录公共检索系统	481
13.1.3 北京大学图书馆公共查询系统	482
13.1.4 清华大学图书馆馆藏目录检索系统	483
13.2 典型中文数字图书检索——超星数字图书馆	486
13.3 典型中文学术期刊论文检索	495
13.3.1 CNKI 中国学术期刊网检索	496
13.3.2 维普中文科技期刊数据库检索	499
13.4 典型外文电子图书检索系统	502
13.4.1 CADAL 外文图书检索	502
13.4.2 世界电子图书馆检索	502
13.4.3 ebrary(电子图书馆)检索	504
13.4.4 OCLC FirstSearch 检索	506
13.4.5 其他典型外文电子图书检索系统简述	508
13.5 典型外文学术期刊检索系统	510
13.5.1 Web of Science 数据库检索	510
13.5.2 IEL 数据库检索	513
13.5.3 EBSCO 学术资源平台检索	518
13.5.4 Wiley 在线图书馆检索	518
13.5.5 其他典型期刊学术论文检索系统	520
本章小结	525
本章思考与练习题	526
参考文献	527

第一部分

信息检索素养基础知识篇

信息检索素养可以描述为：善于根据问题分析自身的信息需求（例如学习或工作需要），进而确定信息来源并使用有效的检索或查找方法，及时地获取需要的信息；善于整理信息、分析评价信息，善于运用信息技术处理信息并用于解决问题；在信息的获取、处理、共享、使用的过程中具有良好的信息意识、信息道德和强烈的社会责任心，有一定创新、协作和服务精神。信息检索意识、信息检索技能和信息利用伦理道德是个体内在信息检索素养的外在表现，也是信息检索素养的基本要素。

第1章说明了信息检索素养的概念含义、发展动因、特点、主要内容与评价标准。同时说明了我国当代大学生信息检索素养的现状，阐述了进行信息检索素养教育与培养的必要性与作用。

第2章阐述了信息检索与知识产权，同时说明了知识产权的含义与内容。本章重点阐述了信息检索与利用的相关法律制度、信息检索与利用过程中的道德自律以及信息检索与利用同知识产权保护相互影响的相互影响。通过本章的学习，旨在培养大学生的信息检索道德和信息获取的相关法律知识。

第3章阐述了信息检索基本知识。包括检索的概念、信息检索的含义与类型、信息检索涉及的相关支撑领域、信息检索的前沿与热点问题。通过本章学习，旨在使读者总体把握信息检索的基本知识。